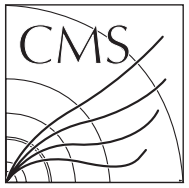
Available on CMS information server

CMS CR -2013/414



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

14 November 2013 (v3, 07 October 2014)

Data preparation for the Compact Muon Solenoid experiment

Giovanni Franzoni for the CMS Collaboration

Abstract

During the first 3 years of operation at the Large Hadron Collider, the Compact Muon Solenoid (CMS) experiment has collected data across evolving conditions of center of mass energy, instantaneous luminosity and collisions pile up. Following this evolution, the CMS collaboration has constantly strived to guarantee the prompt availability of high quality reconstructed data, in order to ensure early and sound physics results. This has relied on a few key areas of constant attention covering careful preparation and maintenance of the event simulation and reconstruction algorithms; efficient and robust strategies and algorithms for the calibration and the alignment of the detector elements; continuous scrutiny of the data quality and the validation of any changes to the software or calibrations which were introduced during the operations. This contribution covers the major development and operational aspects of the CMS offline workflows during the 2010-2013 data taking, underlying its essential role towards the main physics achievements and discoveries of the CMS experiment.

Presented at *IEEE-NSS-MIC-RTDS-2013 2013 IEEE Nuclear Science Symposium and Medical Imaging Conference*

Data Preparation for the Compact Muon Solenoid Experiment

G. Franzoni - CERN, for the Compact Muon Solenoid collaboration

***Abstract*—During the first 3 years of operations at the Large Hadron Collider, the Compact Muon Solenoid detector has collected data across vastly evolving conditions for the center of mass energy, the instantaneous luminosity and the events' pile up. The CMS collaboration has followed this evolution in a continuous way providing high-quality and prompt data reconstruction, necessary for achieving excellent physics performance requested by such high energy physics experiment. The scientific success of CMS came from keeping a constant attention on the key areas: a careful preparation and maintenance of the reconstruction algorithms and core software infrastructure, efficient and robust strategies and algorithms for the calibration and the alignment of the diverse detector elements, up to a continuous and meticulous scrutiny of the data quality and validation of any software infrastructure and detector calibration changes which was deemed necessary. This contribution covers the major development and operational aspects of the CMS offline workflows during data taking experience of 2010-2013, underlying their essential role towards the main physics achievements and discoveries of the CMS experiment during the last years.**

I. Introduction

THIS paper is structured as follows: after a brief introduction describing the Compact Muon Solenoid (CMS) experiment, the overall structure of the data preparation workflows are described. Further details are then given of four key areas of the data preparation process: the measurement and handling of the alignment and calibration constants; the management of data processing and simulated events production; the procedures for data quality assessment and data certification; and the strategies and tools for physics validation.

II. Compact Muon Solenoid

The Compact Muon Solenoid experiment [2] is an omnipurpose detector operating at the Large Hadron Collider [1] at CERN. The central feature of the CMS apparatus is a superconducting solenoid of $6\,\mathrm{m}$ internal diameter, providing a magnetic field of $3.8\,\mathrm{T}$. Within the superconducting solenoid volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass/scintillator hadron calorimeter. Muons are measured in gas-ionization detectors embedded in the steel flux return yoke outside the solenoid. In addition, the CMS detector has extensive forward calorimetry. The first level of the CMS trigger system, composed of custom hardware processors, uses information from the calorimeters and muon detectors to select the most interesting events in a fixed time interval of less than $4\,\mu\mathrm{s}$. The High Level Trigger (HLT) processor farm further decreases the event rate from around 100 kHz to around 0.5 kHz, before data storage.

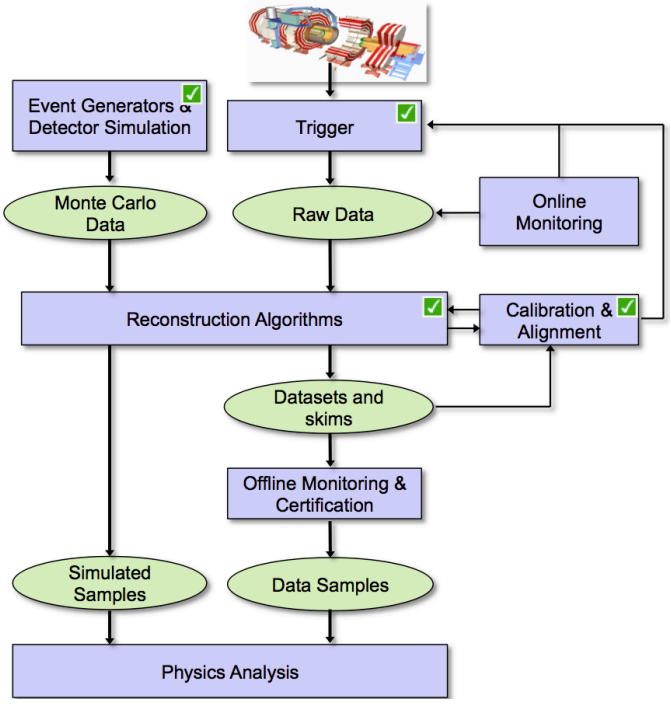


Fig. 1. Workflows devised to provide reconstructed data and simulated data for physics analysis at CMS. The green ticks indicate the steps which are routinely verified and validated.

III. Data Preparation Workflows

For a full exploitation of the collisions delivered by LHC and of the potential of the CMS detector, a complex set of workflows has been devised to provide reconstructed data and simulated data for physics analysis, as shown in figure 1.

The stream of physics events accepted by the CMS trigger system are monitored online during the data taking using a set of distributions made available to the shift crew operating the detector. The raw data acquired are shipped from the CMS to the CERN site, where they are processed to turn the digital output of the readout electronics into reconstructed quantities with a direct physics interpretation such as energies, positions and particle candidates. The reconstruction of the events uses calibration and alignment constants for all the sensitive elements of the detector as a way of incorporating the best knowledge of the detector conditions thus assuring accuracy of the resulting physics quantities. After reconstruction, the events are organised in datasets, defined by a set of outcomes of the HLT selections: this allows aggregating events with common topologies, and eases the distribution of the data for analysis across the CMS computing infrastructure [3]. The

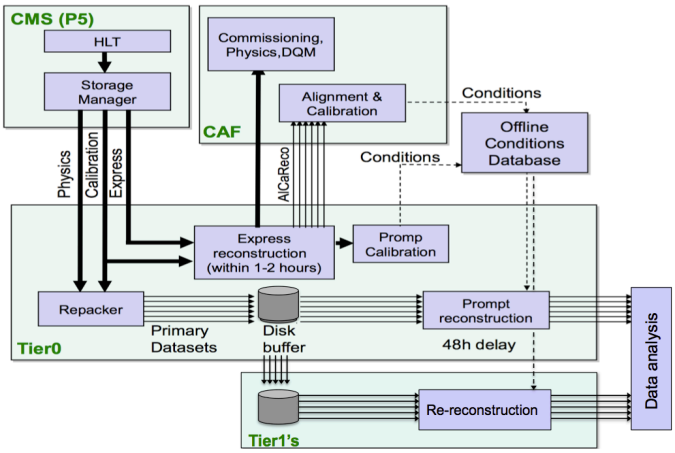


Fig. 2. Workflows for the measurement of alignment and calibration constants and their usage in event reconstruction.

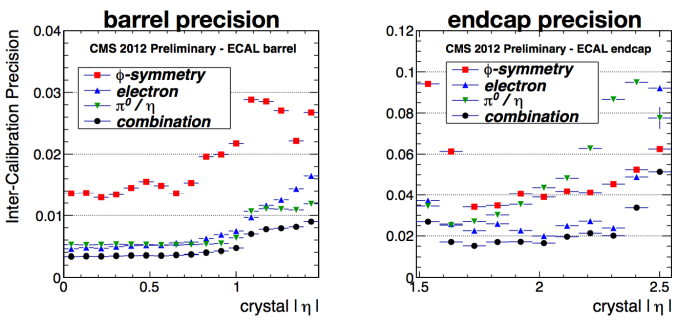


Fig. 3. Relative precision of the channel-to-channel inter-calibrations of the electromagnetic calorimeter.

dataset definition is optimized in order to meet the user needs while minimizing the event duplication.

The production of simulated events stems from the output of physics event generators and emulates the interaction of the final state particles particles with the detector, either with a parameterised [10] or with a detailed simulation based on the GEANT4 toolkit [4]. The reconstruction of the simulated events proceeds along the same lines as the real data, with calibration and alignment constants devised to reflect the accuracy of the conditions used for the real data. In order to match the content of simulated samples to datasets collected with varying experimental conditions (e.g. number of overlaid collisions per bunch crossing, or ageing-related detector properties), a selected subset of the simulated events has been digitised and reconstructed assigning the events to a run number and using run-dependent alignment and calibration sets to emulate the varying experimental conditions.

The complexity and vastness of the infrastructure described requires a thorough set of procedures to verify the correctness and validate the accuracy of all the steps involved. The software of the HLT, of the event reconstruction, and of the event generators and detector simulation are routinely scrutinised, and so are the calibration and alignment constants deployed in production.

IV. Alignment and Calibration Constants

Two hundred alignment and calibration sets are used in the reconstruction of CMS events [5] [6] , covering: channel status calibrations, pedestals and gains for signals amplitude reconstruction, energy scale calibrations, pedestals and drift-times for the drift tubes, Lorentz angle for the silicon detectors and tracker alignment and cross-alignment of the other sub-detectors. Alignment and calibration conditions are handled by three databases, all based on Oracle: one used to serve the data acquisition and provide the detectors' configuration, another to configure the HLT and a third one to deliver conditions to the centralised data production workflows and to the physicists performing analysis.

The procedures for measuring alignment and calibration constants can be categorised according to the available time to complete them. As shown in figure 2, three different classes of alignment and calibration production workflows can be identified:

1 *quasi online*: needed to operate the HLT and for processing of express data stream [3] which reconstructs 10 Hz worth of data right after collection; a key example of these workflows is the measurement of the LHC luminous region position.
2 *prompt calibration loop*: which needs to be completed within 48 h of the data collection; it feeds the *prompt reconstruction* processing performed at CERN to provide reconstructed events for distribution to the CMS collaboration; the *prompt calibration loop* relies on the output of express data streams. The ECAL crystal transparency corrections and the problematic cells flags are key examples of calibration sets produced by the *prompt calibration loop*.
3 *offline for re-reconstruction*: which are used in re-processing campaigns when datasets are reconstructed with all updates and aim at the best possible performance. Predefined dataset with dedicated selection and event contents (see [5] for more details) are used to devise these alignment and calibration constants and optimise the statistics and usage of the resources. The full treatment of detectors alignment inter-dependencies is accounted for in this context, and used for even reconstruction with of all the other alignment and calibration constants updated with respect to *prompt reconstruction*.

Several of the CMS physics results have fundamentally relied on the physics performance obtained by the ultimate accuracy of the alignment and calibration constants; notable among those, the recent observation the Higgs boson with mass near 125 GeV [7]. Because the intrinsic width of the Higgs boson is negligible with respect to its mass, the experimental resolution on the mass, as reconstructed from the decay products, determines the sensitivity for the particle observation. The channel-to-channel inter-calibrations of the electromagnetic calorimeter can be considered as an example to illustrate the impact of calibrations on the physics performance; ECAL inter-calibrations have been measured by combining multiple in-situ techniques [8] with a resulting accuracy shown in figure 3 of less than 1% throughout the barrel, and between 2 and 3% in the endcaps. Combined with the continuous monitoring of

crystal transparency, the accuracy of the ECAL calibration has allowed to achieve an energy resolution for photons ranging from 1.1% to 2.6% in the barrel and from 2.2% to 5% in the endcaps. This resolution, together with the efficiency in determining the vertex of the hard interaction, has been the key ingredient for the Higgs boson observation in the di-photon final state. More details on the impact of alignment and calibration constants on the observation sensitivity in other final states can be found in the reference [7].

V. Data processing and Simulated Events Production

Ten billion proton-proton collision events have been recorded by CMS at the LHC during 2011 ($6.13\,\mathrm{fb}^{-1}$ at $7\,\mathrm{TeV}$) and 2012 ($21.79\,\mathrm{fb}^{-1}$ at $8\,\mathrm{TeV}$) [9], see figure 4. When transferred to the CERN Tier-0 [3] and prior to reconstruction, the events are grouped in non-exclusive datasets according to the result of the HLT selections; 460 datasets have been introduced in total, allowing the efficient distribution of the reconstructed data according to event topologies. Most of the datasets have been reconstructed for a first time by *prompt reconstruction* within a few days of the data taking. In 2012, some dedicated datasets corresponding to the 26% of the total events, have been recorded by CMS but only reconstructed at a later stage, after the end of the physics run. The extra acquired events have allowed the extension of the kinematical phase space available for analysis, without loading the *prompt reconstruction* infrastructure beyond its capacity. Events reprocessing campaigns have taken place on two dozen occasions, most involving a only subset of the datasets, in order to capitalise on improved reconstruction algorithms or calibrations or in order to solve problems in event reconstruction. All the 2011 and 2012 physics dataset have been reprocessed in a final reconstruction campaign, which employes the latest-and-greatest version of the CMS software and of the alignment and calibration constants, and are being used for the publication of legacy physics results.

In addition, twelve billion proton-proton collision events have been simulated at centre of mass energy of 8, 7, 13 and $14\,\mathrm{TeV}$ (the latter two scenarios in the context of studies of the CMS detector upgrade). Physics event generators are used to produce the four-momenta of particles in the final state of a given physics processes; a choice of centre of mass energy, physics generator and physics process define a simulated dataset. After hadronization, the interaction of such particles with the detector is emulated either with a parameterised or with a detailed simulation based on the `GEANT4` toolkit [4]. The response of the sensitive elements to the deposited energies and the digitisation of the signal by the readout electronics are also emulated, to result in the same raw format as the real collision events. The reconstruction of the simulated events proceeds along the same lines and relies on the same software used for the real data, with calibration and alignment constants devised to reflect the accuracy of the conditions used for the real data. All the dataset of generated and simulated events have been reprocessed multiple times by re-executing the digitisation and reconstruction steps with updated calibration and alignment constants, for a total of approximately sixty reprocessing campaigns (most of which involving only a fraction of the simulated datasets).

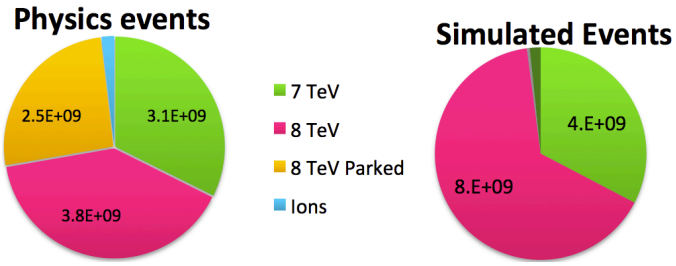


Fig. 4. Statistics of the events collected by the CMS detector (left) and simulated (right).

Managing in a timely and efficient fashion the production of a few thousands simulated datasets per year is critical for the overall efficiency of the scientific production at CMS; it implies aggregating the needs of about twenty active physics groups, establishing priorities among the request and deadlines for their production, and delivering to the computing infrastructure the software configuration needed for the production of the samples. A new web-based platform, McM [11], has been developed and deployed for usage at CMS to manage the production of simulated datasets. Such platform foresees the inter-play between three types of user: the production contact (who reports the needs of a specific physics group and the configurations of the desired samples), the production operators (who assist and troubleshoot the process of testing and submitting to the computing infrastructures the production request), and the production managers (who oversee the overall process and consult with the production operators to establish the relative priorities according, to the needs of the physics community). Besides aggregating information from the three roles, McM handles and triggers all the operations which can be automatised such as submissions of preproduction test jobs, transitions between different production steps and the delivery of notifications to relevant parties about such transitions.

VI. Data Quality and Certification

CMS uses a data quality monitoring (DQM) software framework [12] which provides tools for booking, filling, handling and archiving histograms and which is integral part of the CMSSW event processing software. Such framework is designed around the monitor element object: a wrapper around a `ROOT` [16] histogram class instrumented with quality flags, and string to indicate directory paths. The strength of this data quality framework is that it manages to serve monitoring information in different area of the data preparation at CMS: it provides online monitoring information about the data quality and detector status during the data taking; it collects key monitoring plots filled during the production of small samples dedicated to verify the correctness and validating the physics content of software infrastructure (for event simulation or reconstruction) and of alignment and calibration constants intended for updates; it collects offline key monitoring plots from the production workflows of event simulation or data reconstruction.

TABLE I
NUMBER OF CMS DATASET THAT HAVE PRODUCED AND SENT TO THE DQM GUI A SET OF MONITORING HISTOGRAMS.

Histogram Source	Number of Samples in DQM GUI
Production of Simulated Events	12k
Data reconstruction validation	18k
Validation of Simulated Events	7k

The histograms are served to the users via a web-based graphical user interface (DQM GUI) [13], which provides indexing and historical archival of all the histograms ever produced and is routinely used throughout the CMS collaboration (the DQM GUI receives hundreds of thousands of HTTP daily requests). Data quality histograms from the online and offline monitoring provide input to data certification procedures which determine the sub-set of the recorded collision events usable for physics analysis. The certification is based on the aggregation of reports from experts each using data quality histograms to scrutinise and vet the performance of a given CMS sub-detector or reconstructed physics object.

The datasets produced to validate either a software release or an update of the alignment and calibration constants send to the DQM GUI a considerable set of diagnostic histograms (a few thousands, depending on the context). Table I reports the number of dataset that have produced and sent to the DQM GUI a set of monitoring histograms.

VII. PHYSICS VALIDATION

CMS has put in place a set of procedures and tools to assure the technical correctness (verification) and the physics soundness (validation) of all the data preparation workflows described in the earlier sections. There are several areas in rapid an continuous evolution which need be accompanied and regularly monitored: CMS software framework, events generators, simulation packages and their configuration, be them based on the `GEANT4` [4], reconstruction algorithms, new calibration and alignment constants, dependencies on operative system, compilers, external software packages

The paradigm in usage for verification and validation revolves around a few main principles:

1) campaign for verification and validation is set up by a coordination area whenever either a development software release is made available or in preparation of a large scale reprocessing of data or simulated events; in the following we'll refer to these as *validation campaigns*;
2) all detector groups and physics groups are charged of delivering a report for a given campaign by concentrating the scrutiny on event topologies and set of variables which are of particular interest for their apparatus or physics interest; approximately a hundred collaborators file reports, assuring that the validation campaigns tap as much as possible into the scrutiny expertise spread throughout the collaboration;
3) validation campaigns be performed frequently, to limit the scrutiny to a restricted and defined set of changes; all the bi-weekly pre-releases of the CMS software are validated. Once-off campaigns are performed before initiating a dataset production or deploying limited changes to the HLT or the *prompt reconstruction*;
4) a suit of approximately one hundred small size dataset, some reconstructing real data, others simulating events, are produced for every campaign. The datasets are chosen effectively probe detector effects and physics performance the serving all the reporting groups. All the samples are produced equipped with a large set of diagnostic plots, defined by the validating group and made available via the DQM GUI of the data quality monitoring framework.

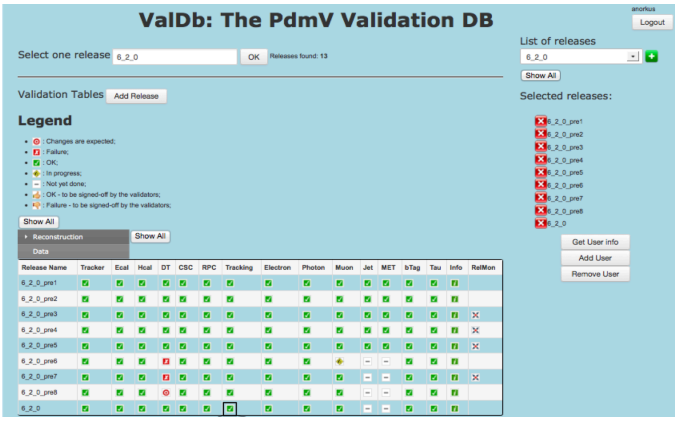


Fig. 5. Screenshot of *ValDb*, where a table reports for different validation campaigns (rows) filled by different groups (columns).

The majority of the validation reports relies on the central and automated workflow described above for the sample and plots production. Over the course of 2011-2013, approximately ten thousand samples have been produced in the context of validation campaigns.

Promptly distributing datasets to the validation experts and aggregating their scrutiny reports has prompted in CMS the development of dedicated web-based tools, which are deployed to ease and economise the overall the verification and validation process. *RelMon* [14] is one such tool, which performs statistical compatibility tests between two instances available in the DQM GUI of the same distribution, one produced from the reference dataset and the other from the target. Distributions are separated in two groups, depending on whether the computed compatibility between target and reference is more or less probable than a given threshold (10^{-5}). *ValDb* [15] is a second tool used to aggregate validation reports for each campaign and each expert involved. Such platform systematises the bookkeeping and is interfaced to an email hyper-news system to distribute the filed validation reports to all the involved experts.

VIII. CONCLUSIONS

The data preparation at CMS has been a crucial ingredient for the full exploitation of proton-proton collisions delivered by the LHC and of the physics performance of the experiment. The prompt availability and ultimate accuracy of alignment and calibrations constants have proven essential for the physics

production of the collaboration, notably for the recent observation of a new boson with mass near 125 GeV. Monitoring the data quality and a continuous effort of verification and validation of software and of the alignment and calibration assure that reconstructed data and simulated samples of excellent quality are fed to physics analyses. The collected experience shows that web-based tools developed to streamline distribution and aggregation of information are a cornerstone of Monte Carlo samples management and physics validation schemes in a large and geographically sparse collaboration like CMS.

ACKNOWLEDGMENT

We congratulate our colleagues in the CERN accelerator departments for the excellent performance of the LHC and thank the technical and administrative staffs at CERN and at other CMS institutes for their contributions to the success of the CMS effort. In addition, we gratefully acknowledge the computing centres and personnel of the Worldwide LHC Computing Grid for delivering so effectively the computing infrastructure essential to our analyses. Finally, we acknowledge the enduring support for the construction and operation of the LHC and the CMS detector provided by the following funding agencies: BMWF and FWF (Austria); FNRS and FWO (Belgium); CNPq, CAPES, FAPERJ, and FAPESP (Brazil); MES (Bulgaria); CERN; CAS, MoST, and NSFC (China); COLCIENCIAS (Colombia); MSES (Croatia); RPF (Cyprus); MoER, SF0690030s09 and ERDF (Estonia); Academy of Finland, MEC, and HIP (Finland); CEA and CNRS/IN2P3 (France); BMBF, DFG, and HGF (Germany); GSRT (Greece); OTKA and NIH (Hungary); DAE and DST (India); IPM (Iran); SFI (Ireland); INFN (Italy); NRF and WCU (Republic of Korea); LAS (Lithuania); CINVESTAV, CONACYT, SEP, and UASLP-FAI (Mexico); MBIE (New Zealand); PAEC (Pakistan); MSHE and NSC (Poland); FCT (Portugal); JINR (Dubna); MON, RosAtom, RAS and RFBR (Russia); MESTD (Serbia); SEIDI and CPAN (Spain); Swiss Funding Agencies (Switzerland); NSC (Taipei); ThEPCenter, IPST, STAR and NSTDA (Thailand); TUBITAK and TAEK (Turkey); NASU (Ukraine); STFC (United Kingdom); DOE and NSF (USA). Individuals have received support from the Marie-Curie programme and the European Research Council and EPLANET (European Union); the Leventis Foundation; the A. P. Sloan Foundation; the Alexander von Humboldt Foundation; the Belgian Federal Science Policy Office; the Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture (FRIA-Belgium); the Agentschap voor Innovatie door Wetenschap en Technologie (IWT-Belgium); the Ministry of Education, Youth and Sports (MEYS) of Czech Republic; the Council of Science and Industrial Research, India; the Compagnia di San Paolo (Torino); the HOMING PLUS programme of Foundation for Polish Science, cofinanced by EU, Regional Development Fund; and the Thalis and Aristeia programmes cofinanced by EU-ESF and the Greek NSRF.

REFERENCES

[1] Lyndon Evans and Philip Bryant, *LHC Machine*, 2008 JINST 3 S08001, doi:10.1088/1748-0221/3/08/S08001.

[2] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[3] D. Bonacorsi, for the CMS Collaboration *Experience with the CMS Computing Model from commissioning to collisions*, 2011 J. Phys.: Conf. Ser. 331 072005, doi:10.1088/1742-6596/331/7/072005.

[4] S. Agostinelli *et al*, *GEANT4: A Simulation toolkit*, Nucl.Instrum.Meth. A506 (2003) 250-303, 072005, doi:10.1016/S0168-9002(03)01368-8.

[5] R. Mankel for the CMS Collaboration *Alignment and calibration experience under LHC data-taking conditions in the CMS experiment*, Journal of Physics: Conference Series 331 (2011) 032022, doi:10.1088/1742-6596/331/3/032022

[6] R. Castello for the CMS Collaboration *Alignment and calibration of CMS detector during collisions at LHC*, The International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), https://indico.cern.ch/contributionDisplay.py?contribId=153\&sessionId=3\&confId=214784

[7] The CMS collaboration *Observation of a new boson with mass near 125 GeV in pp collisions at* $\sqrt{s}$*= 7 and 8 TeV*, JHEP06(2013)081, doi:10.1007/JHEP06(2013)081

[8] The CMS collaboration *Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at*$\sqrt{s}$*= 7 TeV*, JHEP06(2013)081, doi:10.1088/1748-0221/8/09/P09009

[9] The CMS collaboration *Public CMS Luminosity Information*, https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults

[10] S. Abdullin for the CMS collaboration *The Fast Simulation of the CMS Detector at LHC* , Journal of Physics: Conference Series 331 (2011) 032049, doi:10.1088/1742-6596/331/3/032049

[11] J.-R. Vlimant for the CMS collaboration, *MCM : The Evolution of PREP. The CMS tool for Monte-Carlo Request Management*, The International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), http://indico.cern.ch/contributionDisplay.py?contribId=152\&confId=214784

[12] F. De Guio for the CMS collaboration, *The CMS Data Quality Monitoring Software: experience and future prospects*, The International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), http://indico.cern.ch/contributionDisplay.py?contribId=151\&confId=214784

[13] L Tuura for the CMS collaboration, *CMS data quality monitoring web service*, 2010 J. Phys.: Conf. Ser. 219 072055 doi:10.1088/1742-6596/219/7/072055

[14] D. Piparo for the CMS collaboration, *RelMon: A General Approach to QA, Validation and Physics Analysis through Comparison of large Sets of Histograms*, 2012 J. Phys.: Conf. Ser. 396 022011 doi:10.1088/1742-6596/396/2/022011

[15] A. Norkus for the CMS collaboration, *ValDb: an aggregation platform to collect reports on the validation of CMS software and calibrations*, The International Conference on Computing in High Energy and Nuclear Physics (CHEP2013), http://indico.cern.ch/contributionDisplay.py?contribId=149\&confId=214784

[16] *ROOT — A Data Analysis Framework - Cern*, http://root.cern.ch/